Portaalide semantiline koosvõime: põhimõtted ning demosüsteemi ülevaade

Tanel Tammet
detsember 2006

Töö eesmärk:

muuta andmete automaatne kättesaamine (portaalide jaoks ja portaalidest) ja töötlemine võimalikult lihtsaks.

Semantiline koosvõime:

peame silmas andmete (nimed, arvud, kuupäevad, tekstid jne) sisulist kättesaamist edasist töötlemist võimaldaval kujul, mitte lihtsalt nende kuvamist näiteks gif-piltidena

Sissejuhatav kokkuvõte soovitustest

Soovitame esitada andmeid veebilehel selliselt, et:

- andmed oleks harilikul (suvalisel) inimloetaval html-kujul tekstidena. xml formaat ei ole tingimata vajalik.

- harilikul html-lehel olevate andmejuppide ümber pannakse RDFa formaadis metainfo, mis ütleb automaattöötlejale, mis andmetega on tegemist

- lisatav metainfo ei muuda lehe algset väljanägemist ja üldjuhul ei dubleeri lehel juba olevat infot

Näide 1:

Algse html-teksti

```
<tr>
    <td>Jaanus Kask</td>
    <td>6024554</td>
    <td><a href="http://kask.googlepages.com">kliki siia</a></td>
</tr>
```

asemel on html-tekstiks RDFa järgi annoteeritult

```
<tr about="#jaanus_kask">
    <td property="er:nimi">Jaanus Kask</td>
    <td property="er:telefon">6024554</td>
    <td><a rel="er:koduleht"
            href="http://kask.googlepages.com">kliki siia</a></td>
</tr>
```

Näide 2:

Algse html teksti

```
<div>
<a class="tootaja_nimi"
href="http://www.mkm.ee/index.php?id=7187&amp;tootaja=10000450">
Taivo Kivistik</a></div>
<div class="tootaja_inf">
asekantsler (õigusala)
<br>Tel:  6256346 | E-post:
<a class="kontakt_mail" href="javascript:void(0)"
onclick="sendmail('taivo.kivistik', 'mkm.ee')"
>taivo.kivistik
<img src="majandus_juhtkond_files/e_post.gif"
class="e_post_gif" border="0">mkm.ee</a> <br></div><br>
```

Näide 2:

asemel kasutatakse RDFa järgi annoteeritud htmli

```
<span about="#Taivo_Kivistik">
      <div>
      <a class="tootaja_nimi" rel="er:koduleht"
href="http://www.mkm.ee/index.php?id=7187&amp;tootaja=10000450">
      <span property="er:nimi">Taivo Kivistik</span></a></div>
      <div class="tootaja_inf">
      <span property="er:amet">asekantsler (õigusala)</span>
      <br>Tel:  <span property="er:telefon">6256346</span> | E-post:
      <a class="kontakt_mail" href="javascript:void(0)"
      onclick="sendmail('taivo.kivistik', 'mkm.ee')" property="er:email"
      content="taivo.kivistik@mkm.ee">taivo.kivistik
      <img src="majandus_juhtkond_files/e_post.gif"
      class="e_post_gif" border="0">mkm.ee</a> <br></div><br>
</span>
```

Annotatsioonide peamine kasutusstenaarium:

- Asutused annoteerivad oma harilikke veebilehti ise
- Üldportaalid loevad asutuste lehtedelt regulaarselt neid huvitavat infot ja lisavad seda oma andmebaasi
- Üldportaalid kasutavad oma andmebaasi erikuvamiste ja eripäringute jaoks

Taustaks: mis on rdf

Rdf on w3c toetatud universaalne andmeformaat, mis on ette nähtud andmevahetuseks eri süsteemide vahel.

Rdf kujul andmeühikud on neljast osast koosnevad nn andmekolmikud:

- Objekti unikaalne id
- Objekti omaduse nimi
- Objekti omaduse väärtus
- Väärtuse tüüp (number, id, suvaline string, ...)

Näiteks:

“EE3651602310” “nimi” “Tanel Tammet” “string”

HTML ja andmete esitamise põhiküsimus

Reeglina esitatakse andmeid veebis inimloetaval html kujul.

Selline html kuju ei ole kergesti masintöödeldav.

Kas peaks andmeid masintöötluseks esitama eraldi, spetsiaalses formaadis XML-failidena?

Sel juhul oleks meil andmed alati kahes eraldi failis:

- Inimloetav html
- Masintöödeldav xml

Vastus: eraldi xml-formaadis faile ei ole vaja

Täiesti aktsepteeritav variant on mitte-xml-kujul olevas html failis olevate andmete märkimine ehk annoteerimine

- Suvalist html-i saab parsida nagu xml-i
- Muu info lehel ei sega masintöötlemist
- Inimkasutajale annotatsioonid näha ei ole

Annotatsioonide lisamine lehele on väga lihtne

Käsitsi tehtud html-le lisame annotatsioonid käsitsi juurde

Programmiga tehtud html-i puhul lisame annotatsioonide trükkimise programmile juurde

Taustaks: GRDDL, microformats ja RDFa

GRDDL: rdf-andmete võtmine harilikelt veebilehtedelelt
microformats: väikesed spetsiaalsed andmekeeled
RDFa: rdf andmete esitamine annotatsioonidena

Pakutav meetod ongi harilik GRDDL

microformats ja **RDFa** sobivad mõlemad **GRDDL** jaoks

soovitame microformats-i asemel eelistada RDFa-d:

- Universaalsem
- Paindlikum
- W3C soovitus

Linke:

GRDDL	http://www.w3.org/2004/01/rdxh/spec
Microformats	http://microformats.org
RDFa	http://www.w3.org/TR/xhtml-rdfa-primer/

RDFa konkreetsemalt:

Soovitame üldjuhul kasutada ainult neid RDFa võimalusi kui lihtsamaid:

objekti id määramiseks: **about**=”#kohalik_id”
lingi omaduse nime jaoks: **rel**=”er:omadus”
üldiselt omaduse nime jaoks: **property**=”er:omadus”
kui vaja (enamasti ei), tüüp: **type**=”xsd:tyybinimi”
omaduse väärtus: tüüpiliselt htmls olemas
omaduse väärtus alternatiivselt: **content**=”väärtus”

Asutuste lehtede hierarhiad

- Infot kandva lehe lingi juurde panna rel=”er:infoleht” mis ütleb automaattöötlejale, et sellelt lehelt saaks täiendavat (annotatsioonidega) infot

- Lehe algusossa anda lehe enda metainfot, näiteks mis osakonna lehega on tegemist:

Demosüsteemi ülevaade

Demosüsteemis on osaliselt annoteeritud olemasolevad asutuste veebilehed:

- Majandusministeeriumi kontaktinfo üldleht
- Majandusministeeriumi juhtkonna kontaktleht
- Majandusministeeriumi uudiste leht
- Haridusministeeriumi uudiste leht
- Hiiu maavalitsuse kontaktinfo üldleht
- Tartu maavalitsuse kontaktinfo üldleht
- Tartu maavalitsuse ühe töötaja erileht

Mida demosüsteemi lehtedel annoteeriti

- Töötaja nimi, amet, telefon, email jne
- Uudise nimi ja link
- Lehe enda osakond (kui on osakonna leht)
- Linkide kasulikkus automaattöötluseks (kas lingitav leht on asutuse enda leht, mis sisaldab töödeldatavaid andmeid)

Millest demosüsteem koosneb:

- **Andmete agregaator**: loeb avalehelt linke, käib nemad ja edasiviivad lingid läbi, võtab välja RDFa info, salvestab info kõigilt käidud lehtedelt ühte RDF-andmefaili XML kujul

- **Töötajate üld-veebilehe ehitaja:** loeb koostatud universaalset rdf-andmefaili ja teeb töötajate üldise html-lehe, mis sisaldab eri asutuste töötajaid samas formaadis

- **Uudiste üld-veebilehe ehitaja:** loeb koostatud universaalset rdf-andmefaili ja teeb uudiste üldise html-lehe, mis sisaldab eri asutuste uudiseid samas formaadis

Demosüsteemi tehnoloogia

Eeldab, et arvutis on python ja libxml2/libxslt xml-teek

RDF-info eraldab html-st puhta xslt-programmi abil
Töötlus toimub otse pythonis
Andmetabel on lihtne xml-fail
Andmebaasiteeke/mootoreid ei kasutata
Spetsiaalseid (eri)teeke ja mootoreid ei kasutata

Programmide kogusuurused (umbkaudsed) on:

extractrdfa.xsl:	4K
aggregator:	10K
tootajad.cgi:	4K
uudised.cgi:	5K